

dplyr lab

Miriam Schäfer

You can find the github repository for this lab [here](#).

Exercise 1

Question 1

Using distinct (among other functions) compute the number of different songs, artists and musical genre that have been included in the data set. Include the results directly in a presentation text in the markdown document, in the form: the data set contains 584 songs. Notice that the numerical value cannot be copy-pasted from e.g. the console, but has to be included in the text during knitting.

The dataset contains 584 songs from 184 different artists and 50 genres.

Question 2

Compute the number of songs per year and include it in the knitted document as a nicely formatted table (using for instance knitr::kable).

year	Number of songs
2010	51
2011	53
2012	35
2013	71
2014	58
2015	95
2016	80
2017	65
2018	64
2019	31

Question 3

Find the most popular artist in the data set, i.e. the artist with the largest number of songs in the data set. Make sure to count each song only once. Include the name of this artist and the number of songs in the text of the knitted document (as in question 1).

The most popular artist in the dataset is Katy Perry with 17 songs.

Question 4

Compute the minimum, maximum, mean and median bpm as well as the number of songs, for each musical genre. Make sure that each song is used only once in the analysis. Gather the information in a single table included in the knitted result (as in question 2).

top genre	Number of songs	Min BPM	Max BPM	Mean BPM	Median BPM
acoustic pop	2	125	185	155.0000	155.0
alaska indie	1	132	132	132.0000	132.0
alternative r&b	1	136	136	136.0000	136.0
art pop	8	94	150	115.1250	108.5
atl hip hop	4	80	130	108.5000	112.0
australian dance	6	87	192	128.8333	115.0
australian hip hop	1	95	95	95.0000	95.0
australian pop	5	127	186	153.8000	150.0
barbadian pop	15	78	206	125.3333	127.0
baroque pop	2	108	184	146.0000	146.0
belgian edm	2	122	122	122.0000	122.0
big room	10	106	134	122.3000	125.0
boy band	15	77	160	116.1333	120.0
british soul	10	0	164	114.2000	121.5
brostep	2	100	142	121.0000	121.0
canadian contemporary r&b	7	93	186	130.0000	120.0
canadian hip hop	2	100	118	109.0000	109.0
canadian latin	1	98	98	98.0000	98.0
canadian pop	31	65	192	122.1290	119.0
candy pop	2	92	126	109.0000	109.0
celtic rock	1	85	85	85.0000	85.0
chicago rap	1	125	125	125.0000	125.0
colombian pop	3	112	138	126.0000	128.0
complextro	6	102	130	120.3333	127.0
contemporary country	1	86	86	86.0000	86.0
dance pop	324	43	202	118.5309	120.0
danish pop	1	120	120	120.0000	120.0

top genre	Number of songs	Min BPM	Max BPM	Mean BPM	Median BPM
detroit hip hop	2	82	87	84.5000	84.5
downtempo	2	100	167	133.5000	133.5
edm	4	90	118	103.0000	102.0
electro	2	100	116	108.0000	108.0
electro house	1	124	124	124.0000	124.0
electronic trap	2	96	100	98.0000	98.0
electropop	13	90	160	112.8462	103.0
escape room	2	96	158	127.0000	127.0
folk-pop	2	75	86	80.5000	80.5
french indie pop	1	123	123	123.0000	123.0
hip hop	4	80	112	93.7500	91.5
hip pop	6	93	111	100.6667	98.5
hollywood	1	92	92	92.0000	92.0
house	1	122	122	122.0000	122.0
indie pop	2	148	155	151.5000	151.5
irish singer-songwriter	1	129	129	129.0000	129.0
latin	4	94	186	140.7500	141.5
metropopolis	1	127	127	127.0000	127.0
moroccan pop	1	110	110	110.0000	110.0
neo mellow	9	76	138	109.1111	116.0
permanent wave	4	125	148	137.2500	138.0
pop	56	77	190	117.2143	117.0
tropical house	2	120	139	129.5000	129.5

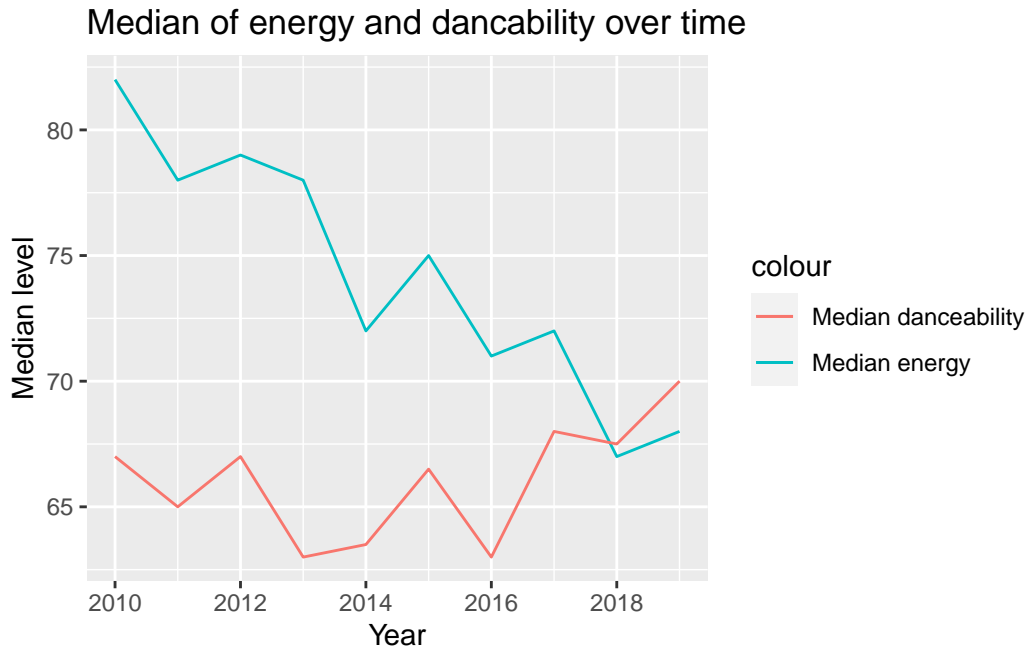
Question 5

Compute the median energy and the median danceability per year in a single data frame.

year	Median_energy	Median_danceability
2010	82	67.0
2011	78	65.0
2012	79	67.0
2013	78	63.0
2014	72	63.5
2015	75	66.5
2016	71	63.0
2017	72	68.0
2018	67	67.5
2019	68	70.0

Question 6

Draw on a single graph the temporal evolution of the median annual energy and the median annual danceability.



Exercise 2

Question 1

Compute the median “Age at enrollment” conditioned both on the Gender and on the “Marital status”.

Gender	Marital status	Median_age_at_enrollment
Male	single	20.0
Male	married	37.0
Male	divorced	40.5
Male	widower	43.0
Male	facto union	34.0
Male	legally separated	55.0
Female	single	19.0
Female	married	34.0
Female	divorced	38.0

Gender	Marital status	Median_age_at_enrollment
Female	widower	21.0
Female	facto union	27.0
Female	legally separated	41.0

Question 2

Transform the data frame obtained in question 1 in order to have three variables: one for the “Marital status”, one for Female and one for Male. Each row should correspond to a specific marital status (given in the corresponding column) while the Female and Male columns should contain the corresponding median age. Include the resulting table in the knitted document as explained in Exercise 1.

Marital status	Male	Female
single	20.0	19
married	37.0	34
divorced	40.5	38
widower	43.0	21
facto union	34.0	27
legally separated	55.0	41

Question 3

Compute the conditional mean of all variables related to “Curricular units” given the value of the Target variable.

Target	Curricular units 1st sem (credited)	Curricular units 1st sem (enrolled)	Curricular units 1st sem
Dropout	0.6094300	5.821253	
Graduate	0.8474423	6.669534	
Enrolled	0.5075567	5.964735	

Question 4

Using the `pivot_` functions, transform the data in order to include in the knitted result a table of the given form.

Table 7: Mean curricular units given the Target variable

Curricular units	Dropout	Graduate	Enrolled
Curricular units 1st sem (credited)	0.61	0.85	0.51
Curricular units 1st sem (enrolled)	5.82	6.67	5.96
Curricular units 1st sem (evaluations)	7.75	8.28	9.34
Curricular units 1st sem (approved)	2.55	6.23	4.32
Curricular units 1st sem (grade)	7.26	12.64	11.13
Curricular units 1st sem (without evaluations)	0.19	0.09	0.18
Curricular units 2nd sem (credited)	0.45	0.67	0.36
Curricular units 2nd sem (enrolled)	5.78	6.63	5.94
Curricular units 2nd sem (evaluations)	7.17	8.14	9.44
Curricular units 2nd sem (approved)	1.94	6.18	4.06
Curricular units 2nd sem (grade)	5.90	12.70	11.12
Curricular units 2nd sem (without evaluations)	0.24	0.08	0.19